Electronic Letters on Computer Vision and Image Analysis 14(3):16-18, 2015

Learning of invariant object recognition in hierarchical neural networks using temporal continuity

Markus Lessmann

Institute for Neural Computation, Ruhr-University Bochum, Germany
Advisor: Rolf P. Würtz, Institute for Neural Computation, Ruhr-University Bochum, Germany
Date and location of PhD thesis defense: 3 November 2014, Ruhr-University Bochum, Germany

Received 25th February 2015; accepted 27th July 2015

Abstract

There has been a lot of progress in the field of invariant object recognition/categorization in the last decade with several methods trying to mimic functioning of the human visual system (e.g. Neocognitron, HMAX, VisNet). Examining those brain regions is a very difficult task with myriads of details to be considered. To simplify modeling approaches, Jeff Hawkins [1] proposed a framework of three basic principles that might underlie computations in regions of the neocortex. These also form the basis for a capable object recognition system named "Hierarchical Temporal Memory" (HTM).

1. Learning of temporal sequences for creating invariance to transformations contained in the training data.
2. Learning in a hierarchical structure, in which lower level knowledge can be reused in higher level context and thereby makes memory usage efficient.
3. Prediction of future signals for disambiguation of noisy input by feedback.

In my thesis I have developed and efficiently implemented two related artificial neural systems relying on these principles, the *Temporal Correlation Graph (TCG)* and the *Temporal Correlation Net (TCN)*. Both are hierarchical neural networks made up of alternating levels of *spatial* and *temporal* neurons located at subsampled image positions called *nodes*. Spatial neurons represent spatial patterns, which on the lowest level are visual features from training images and on higher levels composed patterns of activities. Temporal neurons represent but groups of spatial patterns that tend to follow each other in time. Neural activities are stored in nodes which define the architecture of the systems. In each node any neuron of the same level can become active. Connections from temporal to the next higher spatial level are convergent collecting input from 3×3 spatial neurons in one node. Convergence is chosen thus that at the top of the network only one node of temporal neurons remains. These neurons represent the different object categories the system has learned.

During training each of both systems observes sequences of images showing objects of different categories undergoing transformations in viewing conditions (scaling, rotation in depth, illumination changes etc.) to which the top level responses shall become invariant. First spatial patterns on the lowest level are learned, then

Correspondence to: <markus.lessmann@ini.rub.de>

Recommended for acceptance by Jorge Bernal
DOI http://dx.doi.org/10.5565/rev/elcvia.719
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

temporal groupings in which these tend to appear. Temporal groups (neurons) becoming active concurrently at adjacent positions constitute spatial patterns at the next level and so forth. Thus more and more complex patterns covering observations of increasing time ranges and parts of the input images are created. After learning the systems are able to recognize known object categories e.g. from unknown viewing angles or under unknown illumination conditions.

The systems differ significantly in how spatial and temporal patterns are learned. TCG is trained in batch mode and level by level, for each level all training images are browsed. First, spatial patterns on the lowest level are learned by employing vector quantization on visual features extracted from the images. Thus a codebook or Bag of Words is built containing a descriptive subset of observed patterns. In the next sweep over the training images the occurrences of spatial patterns at the same network position (node) within a short time range are counted. These counts are stored in a matrix which is subsequently clustered to yield temporal groups. Similarity measures are defined that allow to compare spatial patterns of the next level composed of these groups. Thus a codebook of higher level spatial patterns can be learned in the next epoch and again temporal groups can be built. This is continued until neurons on the highest level have been created. Using similarities between temporal groups synaptic connection weights are defined that are used for computing neural activities on test images.

In TCN both learning steps are performed by neural learning rules. The learning of spatial patterns is done by an associative Hebbian learning rule on all but on the lowest level, which still consists of a codebook of visual features. Temporal groups are learned by applying the Trace rule, a Hebbian like learning rule that is also used in VisNet and allows to group neurons that are frequently activated consecutively. No clustering beyond the lowest level is necessary, therefore TCN can be trained in online mode on all levels simultaneously. Every few epochs new neurons are created if current input patterns are represented too weakly.

Whereas TCG is similar to HTM in its learning algorithm it shows better scaling with regard to the number of categories to be learned. This is possible by efficient calculation of neural activities. The learning algorithm of TCN leads to a sparser connectivity structure, which allows for even faster calculation and also could improve generalization in several tests. Both systems were tested on the following databases for object recognition: ETH80, COIL100, ALOI1000, and the German Traffic Sign Recognition Benchmark (GTSRB). On the first 3 databases it was tested how good TCG and TCN can recognize known object categories from unknown viewing angles using different sizes of training and test set. Additionally test conditions were made more difficult by scaling objects, adding structured backgrounds to training and test images and inserting distracting objects that occluded parts of the main objects. On ALOI1000 also tests for recognition under unknown illumination conditions were performed. On ETH80 a leave-one-object-out cross validation test was performed, in which unknown objects of known categories had to be recognized. GTSRB was tested with sorted test images with and without feedback (contradicting benchmark rules). Both systems performed very well in the standard tests even with very small training sets, in most tests TCN outperformed TCG with reaching for example over 90% recognition when learning from ca. 6% of all ETH80 images and close to 100% for bigger training sets. With adverse training conditions both systems decreased in performance depending on the applied manipulation. Occlusion and structured background had moderate effects with up to ca. 20% lower recognition rates, scaling was much more harmful for scaling factors $\notin [0.9, 1.2]$. Illumination tests performed close to 100%, and recognition rates of 98.33% (hence close to benchmark winners) could be reached in GTSRB using feedback.

References

[1] Jeff Hawkins, “On Intelligence”, Times Books, 2004.

[2] Markus Lessmann and Rolf P. Würtz, “Learning of invariant object recognition in a hierarchical network”, *Neural Networks*, 54:70-84, 2014.

[3] Markus Lessmann, “Learning of invariant object recognition in hierarchical neural networks using temporal continuity”, PhD thesis, Ruhr-University Bochum, 2014